Exploring Genetic and Lifestyle Factors Affecting Diabetes Using Logistic Regression Modelling

Zirui Deng

Faculty of Business, The Hong Kong Polytechnic University, Hong Kong. 999077, China

22102565D@connect.polyu.hk

Abstract. Diabetes mellitus is a globally widespread chronic disease with a rising prevalence and a huge impact on individual health, healthcare systems and socio-economics. Studying the factors influencing diabetes can help identify at-risk patients, prevent or delay onset, and optimize personalized treatment. The aetiology of diabetes is complex, and Existing studies have shown that diabetes occurs and develops as a result of a combination of factors, including genetic, lifestyle and environmental factors. In this study, several factors that may influence diabetes were systematically analysed using logistic regression modelling. Five factors were selected from 15 possible influences (including genetic and lifestyle factors) that had a significant effect. They are family history of diabetes, diet quality, fasting blood glucose concentration, thirst level, and hypertension. Hence, having a family history of diabetes, poorer diet quality, higher fasting blood glucose concentration, greater thirst, and higher blood pressure are more likely to develop diabetes. Such high-risk groups should make lifestyle modifications to reduce the risk of diabetes through a healthy diet, regular work and rest, and moderate exercise.

Keywords: Diabetes, logistic regression modelling, genetic, lifestyle factors.

1. Introduction

According to the International Diabetes Federation, the number of people with diabetes worldwide has exceeded 400 million and is expected to continue to grow in the coming decades [1]. The number of people with diabetes is growing so rapidly that diabetes has become a significant public health problem worldwide. This growth not only affects the health of individuals but also poses a considerable challenge to public health systems.

Diabetes mellitus is a chronic metabolic disease caused by insufficient secretion or defective insulin action, manifesting as a prolonged increase in blood glucose [2]. Moreover, the aetiology of diabetes is complex and influenced by various genetic and lifestyle factors. Family genetic factors are important risk factors for diabetes [3]. Unhealthy lifestyles such as high-sugar and high-fat diets, lack of exercise, and smoking are important lifestyle factors affecting diabetes [4]. However, there are fewer multivariate analyses combining genetic and lifestyle factors to explore the impact on diabetes, especially combined studies combining physiological indicators, symptomatic manifestations, and quality of life scores, which remains an important research gap.

The study of factors influencing diabetes has important public health implications. Firstly, by identifying people at risk, effective interventions such as lifestyle changes can be made before the onset of diabetes, thus preventing or delaying its onset. This not only helps to reduce the overall incidence of diabetes but also reduces the burden on the healthcare system (Roglic, 2016). Secondly, understanding an individual's genetic background and lifestyle factors can help to develop a personalized treatment plan and improve treatment outcomes and patient compliance. In addition, the results of the study can provide a scientific basis for governments and health organizations to formulate health policies, such as policies to promote healthy diet and exercise. This helps to rationally allocate medical resources and prioritize support for preventing and treating high-risk groups [5]. Through effective management and intervention, the likelihood of developing diabetes mellitus can be reduced, and the health of patients can be safeguarded.

Many models can be used to predict how much risk factors affect diabetes, with decision trees having the highest classification accuracy, logistic regression having the next highest, and artificial neural networks (ANN) having the lowest [6]. However, logistic regression output can more clearly explain the magnitude of individual characteristics and their positive or negative influence on the results. Therefore, this article aims to explore the main factors influencing diabetes by combining genetic and lifestyle factors in a multivariate analysis. The article will explore the main factors influencing diabetes using logistic regression modelling. From the 15 possible influencing factors (including genetic and lifestyle factors), five main influencing factors will be selected to summarise the characteristics of the high-risk group. Based on these results, specific recommendations for living with diabetes will be made.

2. Methodology

2.1. Data source and description

The dataset was obtained from Kaggle and contains 45 health indicators (including genetic and lifestyle factors) of the patients and their labels of whether they have diabetes or not. The dataset contains comprehensive health data for 1879 patients, including lifestyle factors, medical history, clinical measures, medication use, symptoms, quality of life scores, environmental exposures, and health behaviours.

2.2. Indicator selection and description

In this paper, 15 key health indicators were selected from the original 45 sets of health indicator data. Reis et al. pointed out that dietary intake, BMI, physical activity, and health literacy impact diabetes [7]. Howard et al. stated that total cholesterol levels and blood pressure also affect diabetes [8]. Based on this, 15 indicators related to genetic and lifestyle factors were screened for multivariate analysis in this paper, as shown in Table 1.

Table 1. Data Description

Factors	Unit	Data Format
Age	Years	20-29 years
Gender	/	0 = Male, 1 = Female
BMI	BMI Index	15-40
Weekly Physical Activity Time	Hours	0-10 hours
Diet Quality	Quality Index	0-100 score
Sleep Quality	/	0-10 score
Family History of Diabetes	/	0 = No, 1 = Yes
Fasting Blood Glucose	mg/dL	70-200 mg/dL
Total Cholesterol	mg/dL	100-250 mg/dL
Triglycerides	mg/dL	50-400 mg/dL
Fatigue Level	/	0-10 score
Quality of Life Score	Score (0-10)	0-10 score
Thirst Level	Score (0-10)	0-10 score
Hypertension	/	0 = No, 1 = Yes
Health Literacy	Score (0-10)	0-10 score

2.3. Introduction to the Methodology

This paper aims to explore and analyze the effect of different health indicators on the risk of developing diabetes using a binary logistic regression model. This study will use diabetes mellitus as the dependent variable and various health indicators (e.g., age, body mass index (BMI), blood glucose level, physical activity, dietary habits, etc.) as independent variables. The magnitude of the effect of these indicators on the onset of diabetes is assessed by building and analyzing models.

1) Logistic regression begins with the calculation of a linear combination of the input variables X:

$$z = W_0 + W_1X_1 + W_2X_2 + \cdots + W_nX_n + \varepsilon \tag{1}$$

2) The linear combination z is then converted to a probability value p using the Sigmoid function $\sigma(z)$:

$$p = \sigma(z) = \frac{1}{1+e^{-Z}} \tag{2}$$

3) After that, the threshold method is used for classification. If p≥0.5, Y=1 is predicted. if p<0.5, Y=0 is predicted.

3. Results

3.1. Logistic regression analysis

According to Chart 2, quality of diet, family history of diabetes, fasting blood glucose concentration, thirst level, and high blood pressure had significant effects on diabetes mellitus. Among them, fasting blood glucose concentration has the most significant effect on diabetes. With a correlation coefficient of 1.0878, hypertension is the factor that significantly affects diabetes mellitus. With a value of 0.6060, thirst has the second-most significant impact on diabetes mellitus. With a coefficient of 0.2763, a family history of diabetes mellitus has the third-most significant impact on the condition. Detary quality is the fourth most significant effect, with a coefficient of -0.0401. fasting blood glucose concentration has the fifth most significant effect on diabetes mellitus, with a coefficient of 0.0313. Its coefficient is 0.0313, while the other influences are not significant on diabetes from the data.

Table 2. Correlation coefficients of characteristic variables

Factors	Coefficient	P-value	Marginal Effect
Age	0.0004	0.875	0.0001
Gender	0.0966	0.375	0.0224
BMI	0.0086	0.266	0.0020
Weekly Physical Activity Time	-0.0081	0.672	-0.0019
Diet Quality	-0.0401	0.035***	-0.0093
Sleep Quality	-0.0025	0.938	-0.0006
Family History of Diabetes	0.2763	0.029***	0.0642
Fasting Blood Glucose	0.0313	2.639e-76***	0.0073
Total Cholesterol	-0.0013	0.306	-0.0003
Triglycerides	0.0006	0.246	0.0001
Fatigue Level	0.0003	0.988	0.0001
Quality of Life Score	0.0018	0.336	0.0004
Thirst Level	0.6060	1.349e-5***	0.1408
Hypertension	1.0878	9.890e-13***	0.2528
Health Literacy	-0.0100	0.5947	-0.0023

According to the study's findings, those who have a family history of diabetes, consume a poorer diet, have higher fasting blood glucose levels, are more thirsty, and have higher blood pressure are all at a higher risk of developing the disease. Each of these influencing factors will be briefly analyzed below.

People who have a family genetic background of diabetes are more prone to get diabetes, according to the main influencing element of family history of diabetes. Diabetes is strongly associated with genetic factors. It has been found that specific genetic variants (e.g., genes that affect insulin secretion and insulin sensitivity) increase an individual's risk of developing the disease [9]. Suppose a parent or close relative has a history of diabetes. In that case, the child is more likely to carry the relevant susceptibility genes and thus is more likely to develop diabetes. A negative diet quality coefficient indicates that while an unhealthy diet raises the chance of developing the disease, a healthy diet may help lower it. Diabetes is more prone to develop in people with higher fasting blood glucose levels. High fasting blood glucose usually means insulin resistance or impaired beta cell function, leading to abnormal blood glucose regulation. This increases the risk of diabetes over time and is one of the key signals and diagnostic indicators of prediabetes.

Greater thirst may affect diabetes as it can indicate elevated blood sugar. Hyperglycaemia leads to osmotic diuresis, which causes the body to urinate heavily, leading to dehydration and exacerbating blood glucose imbalance [10]. High blood pressure can affect diabetes. This is because it damages blood vessels and reduces the sensitivity of tissues to insulin, leading to insulin resistance and exacerbating the difficulty of glycaemic control [11].

3.2. Model evaluation

Figure 3 shows that the model's precision (Precision) for category 0 (no diabetes) is 0.77, meaning that 77% of the samples it predicted to have no diabetes are actually not diabetic. With a Recall (Recall) of 0.73, the model accurately predicts that 73% of all non-diabetic samples are not diabetic. The F1 score, which is the sum of precision and recall, is 0.75.

60% of the samples that the model predicted to have diabetes were actually diabetic, according to the model's precision of 0.60 for category 1 (having diabetes). With a recall of 0.64, the model accurately predicts that 64% of all samples with diabetes have the disease. The F1 score, which is the sum of the precision and recall, is 0.62. The model's total precision was 0.70, meaning that 70% of the samples were properly identified. The model's weighted average precision is 0.70, its weighted average recall is 0.70, its F1 score is 0.70, its macro average precision is 0.68, its recall is 0.69, and its F1 score is 0.69. Taken together, although the model can identify people at high risk of diabetes to a certain extent, the model cannot identify people at high risk of diabetes due to the multifactorial nature of the occurrence of diabetes. However, because the occurrence of diabetes mellitus is affected by multiple factors, there are still some limitations in the model's predictive ability.

Table 3. Classification report

	precision	recall	F1-score	support
0	0.77	0.73	0.75	233
1	0.60	0.64	0.62	143
accuracy			0.70	376
Macro avg	0.68	0.69	0.69	376
Weighted avg	0.70	0.70	0.70	376

The overall performance accuracy of the model was 70%, which is not very accurate. The prediction was more accurate for not having diabetes (category 0) but poor for having diabetes (category 1), and further optimization of the model may be required. Other classification algorithms (e.g., Random Forest, XGBoost, etc.) can be used to improve the model's diabetes prediction. Data checking removes the influences of low relevant features and tries feature combinations or normalization to improve the model's accuracy. Therefore, in practical applications, a comprehensive assessment should be conducted by combining multiple health indicators to improve the accuracy of identifying people at risk.

4. Conclusion

Through the study, this paper found that in the multivariate analysis combining genetic and lifestyle factors, the influencing factors that significantly affect diabetes are family history of diabetes, quality of diet, fasting blood glucose concentration, level of thirst, and high blood pressure. The other influencing factors were statistically insignificant in terms of diabetes mellitus. Therefore, people with a family history of diabetes, poorer diet quality, higher fasting blood glucose concentration, greater thirst, and higher blood pressure are more likely to develop diabetes.

Based on the results of the study, people at high risk of diabetes should take the following steps to reduce their risk of developing the disease or improve their condition. Blood glucose levels should be checked often for those with a family history of diabetes in order to identify issues and take prompt action. People should consume more dietary fiber and less high-fat and high-sugar foods. Eat more nutritious food and change unhealthy eating habits. People at high risk avoid high salt diets to reduce the risk of hypertension.

This paper conducted a multivariate analysis using logistic regression modelling and combining genetic and lifestyle factors, especially on the comprehensive consideration of physiological indicators, symptomatic manifestations and quality of life scores. The multivariate interactions were explored outside the framework of univariate analyses or single-factor intra-regional analyses. However, the accuracy of the model used in this paper is limited, and the small sample size may still affect the reliability of the conclusions. Other more advanced classification algorithms can be tried in future studies to improve the model's ability to predict diabetes. Moreover, the influences of low correlation features should be removed, and feature combinations or standardization should be tried to obtain more reliable findings.

References

[1] S. Sen., R. Chakraborty., & B. De. Diabetes mellitus in the 21st century (pp. 55-67). New York, NY, USA: Springer (2016).

[2] M. N. Piero., G. M. Nzaro., & J. M. Njagi. Diabetes mellitus-a devastating metabolic disorder (2015).

[3] K. Aasbjerg., C. H. Nørgaard., N. Vestergaard., P. Søgaard., L. Køber., P. Weeke., ... & C. Torp-Pedersen. Risk of diabetes among related and unrelated family members. Diabetes Research and Clinical Practice, 160, 107997 (2020).

[4] F. B. Hu. Globalization of diabetes: the role of diet, lifestyle, and genes. Diabetes care, 34 (6), 1249–1257 (2011).

[5] J. Hill., M. Nielsen., & M. H. Fox. Understanding the social factors contributing to diabetes: a means to informing health care and social policies for the chronically ill. The Permanente Journal, 17 (2), 67 (2013).

[6] X. H. Meng., Y. X. Huang., D. P. Rao., Q. Zhang., & Q. Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29 (2), 93–99 (2013).

[7] J. P. Reis., C. M. Loria., P. D. Sorlie., Y. Park., A. Hollenbeck., & A. Schatzkin. Lifestyle factors and risk for new-onset diabetes: a population-based cohort study. Annals of internal medicine, 155 (5), 292-299 (2011).

[8] B. V. Howard., M. J. Roman., R. B. Devereux., J. L. Fleg., J. M. Galloway., J. A. Henderson., ... & J. Zhu. Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: the SANDS randomized trial. Jama, 299 (14), 1678-1689 (2008).

[9] A. E. Brown., & M. Walker. Genetics of insulin resistance and the metabolic syndrome. Current cardiology reports, 18 (8), 75 (2016).

[10] P. Newland-Jones., M. Patel., & K. Dhatariya. Acute metabolic complications of diabetes: diabetic ketoacidosis and the hyperosmolar hyperglycaemic state in adults. Textbook of Diabetes, 602-611 (2024).

[11] R. Izzo., G. de Simone., M. Chinali., G. Iaccarino., V. Trimarco., F. Rozza., ... & N. De Luca. Insufficient control of blood pressure and incident diabetes. Diabetes care, 32 (5), 845-850 (2009).